Faculty of Engineering and Information Technology
University of Technology Sydney

Text and Data Mining for Human Drug Understanding

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Yi Zheng

October 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yi Zheng declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Advanced Analytics Institute, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

SIGNATURE OF CANDIDATE: _____

[Yi Zheng]

DATE: 13th June, 2019

PLACE: Sydney, Australia

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Jinyan Li for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me a lot in all the time of my research and writing of this thesis. With his help, my research skills such as scientific writing, academic communication, and presentation have been improved to a new stage. Without his guidance and persistent help, this thesis and related research work would not be possible.

I would like to thank my co-supervisor Prof. Longbing Cao for his help on the enrollment of UTS and valuable suggestions for my research work. I also thank my Master's supervisor Quanyuan Wu, for his strong support when I apply for the CSC (China scholarship council) scholarship. Many thanks to Dr. Jie Yin and Dr. Xiaoying Gao, two of my research partners, for their insightful suggestions and patience to discuss the research problems with me.

My sincere thanks also go to my research team members Hui Peng, Zhixun Zhao, Xiaocai Zhang, Chaowang Lan, and Yuansheng Liu for their help in both my research and daily life. I am grateful to three former team members Dr. Jing Ren, Dr. Shameek Ghosh, and Dr. Renhua Song, for their kind help at the beginning of my Ph.D. study. I also want to thank Tao Tang and Xuan Zhang who join us recently. It's my honor to be one member of my research group. Thank you all for bringing me the feeling of home in Australia. I will memorize the unforgettable experiences forever.

In addition, I am grateful to acknowledge the funding sources of my study and conference travel, including the tuition fee and living expenses provided by the China Scholarship Council and Graduate Research School, travel funds provided by Faculty of Engineering and Information Technologies. Thanks to all staffs of Advanced Analytics Institute and School of Software who provide services and conveniences to my study and research in UTS.

At last, I really appreciate my wife Chengcheng Sun and my parents for their strong support during my overseas study. Especially my wife, she sacrificed her time and opportunities to support my study. A thank you to my relatives and friends in China, for your concern about my life and study abroad.

Yi Zheng
June 2019 @ UTS

Contents

Certificate . . . . . **i**
Acknowledgment . . . . . **iii**
List of Figures . . . . . **ix**
List of Tables . . . . . **xv**
List of Publications . . . . . **xvii**
Abstract . . . . . **xxi**

Chapter 1 Introduction **1**
1.1 Background 1
1.1.1 Side-effects, targets, and drug-drug interactions 1
1.1.2 Pharmacologic databases and social media 4
1.1.3 Text mining and data mining 7
1.2 Research topics 10
1.3 Research contributions 13
1.4 Thesis structure 16

Chapter 2 Related work and literature review **19**
2.1 Drug side-effect prediction from pharmacologic databases . . . 19
2.1.1 Drug side-effect prediction for single drug medication . 19
2.1.2 Drug side-effect prediction for combined medication . . 22
2.2 Adverse drug reaction detection from social media 25
2.3 Drug-target association identification 26
2.4 Drug-drug interaction detection 28
2.5 Summary 30

Chapter 3 Inverse similarity and reliable negative samples for drug side-effect prediction 31

3.1 Introduction 31

3.2 Materials 32

3.2.1 Drug side-effect profiles 32

3.2.2 Drug similarities 34

3.3 Methods 36

3.3.1 Drug similarity integration framework 36

3.3.2 Negative sample selection based on comprehensive drug similarities for side-effect predictions 37

3.4 Results and discussions 39

3.4.1 Evaluation on drug similarity integration framework . . 39

3.4.2 Evaluation on balanced and imbalanced training set . . 41

3.4.3 Performance improvement brought by the selection of highly-reliable negative samples 45

3.4.4 Comparison with other methods 48

3.4.5 Drugs that have the predicted side-effect "drug eruption": a case study 49

Chapter 4 Predicting side-effects of combined medication from heterogeneous databases 53

4.1 Introduction 53

4.2 Methods 54

4.2.1 Data resources 54

4.2.2 Proposed method 55

4.3 Results and discussions 61

4.3.1 Parameter optimization 61

4.3.2 Evaluation on classic classifiers 62

4.3.3 Comparison with baseline methods 63

4.3.4 Predicted adverse drug reactions for the drug pair "Albuterol-Zolpidem": a case study 64

Chapter 5 Constrained information entropy for detecting adverse drug reactions from medical forums . . . 68
5.1 Introduction . . . 68
5.2 Methods . . . 69
5.2.1 Framework . . . 69
5.2.2 Data collection . . . 69
5.2.3 Preprocessing . . . 70
5.2.4 Entity extraction . . . 70
5.2.5 ADRs detection . . . 70
5.3 Results and discussions . . . 72
5.3.1 Dataset and evaluation metrics . . . 72
5.3.2 Adverse drug reaction detection . . . 73

Chapter 6 Predicting drug targets using anchor graph hashing and ensemble learning . . . 77
6.1 Introduction . . . 77
6.2 Methods . . . 78
6.2.1 Prediction framework . . . 78
6.2.2 Data representation . . . 80
6.2.3 Anchor graph hashing compression . . . 80
6.2.4 Sample selection . . . 81
6.2.5 Ensemble learning . . . 81
6.2.6 10-fold cross-validation . . . 82
6.3 Results and discussions . . . 82
6.3.1 Data resources . . . 82
6.3.2 Parameter optimization . . . 83
6.3.3 Comparison with existing methods . . . 84

Chapter 7 DDI-PULearn: a novel PU learning method for prediction of drug-drug interaction . . . 89
7.1 Introduction . . . 89
7.2 Methods . . . 90

7.2.1 Data resources 90
7.2.2 Proposed methods 91
7.3 Results and discussions 97
7.3.1 Components for PCA 97
7.3.2 Representation of DDIs using multi-source drug property data 98
7.3.3 Performance improvement brought by identified reliable negative samples 99
7.3.4 Comparison with existing state-of-the-art methods . . . 101
7.3.5 Novel DDIs predicted by DDI-PULearn 103

Chapter 8 Conclusions and future work 105
8.1 Conclusions 105
8.2 Future work 108

Chapter A Appendix: Methodology foundation 111
A.1 Applied machine learning algorithms 111
A.1.1 K-nearest neighbors 111
A.1.2 Support vector machine 111
A.1.3 Extreme learning machine 112
A.1.4 Radial basis function networks 113
A.1.5 Logistic regression 113
A.1.6 Random forest 114
A.1.7 XGBoost 115
A.2 Cross validation and performance evaluation indices 115
A.2.1 Cross validation 115
A.2.2 Performance evaluation indices 116

Chapter B Additional files 117

Chapter C Appendix: List of Symbols 118

Bibliography 121

List of Figures

1.1 **Thesis Structure.** It consists of the following four parts: introduction, related work, my work, and conclusions and future work. Short introduction of each part is shown in the right side. 18

3.1 **Characteristics of side-effects and their associated drugs.** The left panel is the index-plot of the number of associated drugs for each side-effect and the right panel is the histogram of the associated drug number for the side-effects. . 33

3.2 **Flow diagram of drug side-effect prediction with the proposed negative sample selection method using the comprehensive drug similarity.** 38

3.3 **Boxplots of the $F1$-scores for different similarity measurements and different similarity integration methods using the KNN classifier.** 41

3.4 **Differences among similarity measurements and similarity integration methods.** In each panel, the x-axis denotes the index of each side-effect and the y-axis denotes the $F1$-score difference between two methods. For instance, Figure 3.4 (a) describes the differences between "ComMean" and "ComGM" using the $F1$-score of each side-effect from ComMean minus that from ComGM (i.e. difference=$F1$-score(ComMean)$-F1$-score(ComGM)). Thus we can identify which method performs better by comparing the area under the curve above zero (i.e., area A) with the area above the curve under zero (i.e., area B). 42

3.5 **Scatter plots of $F1$-scores for different classifiers using the comprehensive similarity on balanced and imbalanced training sets.** The x-axis denotes $F1$-scores of results based on imbalanced training sets, and the y-axis for balanced training sets. The line "$y = x$" on which $F1$-scores are equal, is the reference line to better visualize the results. Dots above the reference line are colored green while dots below the reference line are colored red. 44

3.6 **Comparison results using the proposed negative sample selection method and random sample selection method.** The x-axis denotes $F1$-scores of results based on negative samples selected randomly, and the y-axis denotes $F1$-scores of results based on negative samples selected by the proposed method. The line "$y = x$" on which $F1$-scores are equal, is used as the reference line. Dots above the reference line are colored green while dots below the reference line are colored red. 46

3.7 **The top 50 drugs which are predicted to have the side-effect "drug eruption".** Labels on the edges illustrate the ranks of predicted associations and the confirmation types. The symbols "#" and "$" denote that the corresponding associations can be validated by records from the side-effect database SIDER (colored green) and related literature (colored red) respectively. The symbol "?" means the predicted associations cannot be validated to the best of our knowledge (colored orange). 51

4.1 **The framework of HCNS.** It consists of three components: drug representation, credible negative sample generation, and drug-drug-side-effect association prediction. 57

4.2 **The macro-averaging *F*1-scores with different PCA component numbers and different negative sample ratios.** 62

4.3 **The macro-averaging precision, recall, *F*1-score and accuracy of HCNS and other three comparison methods.** 65

4.4 **The top 40 side-effects which are predicted to be associated with the drug pair "Albuterol-Zolpidem".** Labels on the edges illustrate the ranks of predicted associations and the confirmation types. "#" denotes the relation is known in the Tatonetti Lab dataset, "$" means the relation is the common side-effects of the drug pair, "?" indicates there are no evidence for the relation. 66

5.1 **Framework of CIE.** It consists of four components namely data collection, preprocessing, entity extraction and ADRs detection. 69

5.2 **ADRs detection performance from SteadyHealth.Com using the co-occurrence based method with and without the keyword filter**. 74

6.1 **The framework for drug target prediction by anchor graph hashing and ensemble learning.** It consists of five components: data representation, anchor graph hashing compression, sample selection, ensemble learning, and 10-fold cross-validation. 79

6.2 **Characteristics of targets and their associated drugs.** The left panel is the histogram of the associated drug number for the targets and the right panel is the index-plot of the number of associated drugs for each target. 83

6.3 **The AUC of AGHEL with different AGH settings.** The x-axis is the number of output hashing bits and the y-axis is the AUC score. 84

6.4 **The average execution time (s) of four methods using 10-fold cross-validation.** 85

6.5 **The ROC curves of four methods using 25 runs of 10-fold cross-validation.** 86

6.6 **The average AUC of four methods evaluated by 25 runs of 10-fold cross-validation.** 87

7.1 **The framework of the proposed method.** It consists of the following five components: reliable negative sample identification, feature vector representation for DDIs, PCA compression, DDI prediction, and performance evaluation. RN: reliable negative samples; PCA: principal component analysis; DDI: drug-drug interaction. 92

7.2 **The flow chart for the identification of reliable negative samples.** OCSVM: one-class support vector machine; KNN: k-nearest neighbor; RNS: reliable negative samples; RU: remaining unlabeled. 94

7.3 ***F*1-scores of DDI-PULearn with different PCNs.** The x-axis is the PCA component number and the y-axis is the F1-score. Panel (a) shows the F1-scores for PCN between 1 and 2,000, and Panel (b) is an amplification of the range [20,150]. . 98

7.4 **Prediction results using different combinations of drug features.** BDPs refer to the basic drug properties namely drug chemical substructures, drug targets, and drug indications. 100

A.1 **The structure of RBF.** 114

List of Tables

1.1 **Pharmacologic databases used in this thesis.** 5

3.1 **The drug side-effect dataset.** 33

3.2 **Macro-averaging $F1$-score, precision, and recall of four typical classifiers based on negative samples selected by the proposed negative sample selection method and randomly selected negative samples.** . . . 48

3.3 **Performance of the proposed method and state-of-the-art-methods using 5-fold cross-validation on Liu's data set.** 50

4.1 **Macro-averaging AUC, $F1$-score, precision, recall and accuracy of four typical classifiers based on negative samples selected by HCNS and RGNS.** 63

5.1 **A set of keywords or phrases denote the causality relationship between drugs and ADRs.** 71

5.2 **Dataset summary of SteadyHealth.Com** 73

5.3 **Dataset summary of MedHelp.Org** 73

5.4 **Top 20 identified ADRs from SteadyHealth.Com** 75

5.5 **Top 20 identified ADRs from Medhelp.Org** 76

6.1 **Dataset used in this chapter and two datasets used in publications.** 83

6.2 **Average metric scores of four methods evaluated by 25 runs of 10-fold cross-validation.** 87

7.1 **Prediction performance comparison with the two baseline methods, namely all-negatives and random-negatives.** . 101
7.2 **Performances of DDI-PULearn and the benchmark methods evaluated by 20 runs of 3-fold cross-validation.** 102
7.3 **Performances of DDI-PULearn and the benchmark methods evaluated by 20 runs of 5-fold cross-validation.** 102
7.4 **Top 25 novel DDIs predicted by the proposed method DDI-PULearn.** (DDIs which are confirmed in DrugBank are highlighted in bold font.) 104

List of Publications

The journal and conference papers published during my PhD study are listed as follows:

Related to the Thesis :

1. **Y. Zheng**, H. Peng, J. Li, et al. Inverse similarity and reliable negative samples for drug side-effect prediction [J], BMC Bioinformatics, 2019, 19(13): 554.

2. **Y. Zheng**, H. Peng, J. Li, et al. Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases [J], BMC Bioinformatics, 2018, 19(S19).

3. **Y. Zheng**, H. Peng, J. Li, et al. Predicting Drug Targets from Heterogeneous Spaces using Anchor Graph Hashing and Ensemble Learning [C], 2018 International Joint Conference on Neural Networks. IEEE, 2018: 1-7.

4. **Y. Zheng**, S. Ghosh, J. Li, et al. An Optimized Drug Similarity Framework for Side-effect Prediction [C], Computing in Cardiology. IEEE, 2017: 1-4.

5. **Y. Zheng**, C. Lan, H. Peng, J. Li, et al. Using constrained information entropy to detect rare adverse drug reactions from medical forums [C], 2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC). IEEE, 2016: 2460-2463.

6. **Y. Zheng**, H. Peng, J. Li, et al. Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces [J], The 30th International Conference on Genome Informatics 2019 (transferred to BMC Bioinformatics). **(conference accepted, journal potentially accepted)**.

7. **Y. Zheng**, H. Peng, J. Li, et al. DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions [J], International Conference on Bioinformatics 2019 (transferred to BMC Bioinformatics). **(conference accepted, journal potentially accepted)**.

Others :

8. **Y. Zheng**, S. Ghosh, T. Lammers, J. Li, et al. Deriving Public Sector Workforce Insights: A Case Study using Australian Public Sector Employment Profiles [C], 12th International Conference on Advanced Data Mining and Applications (ADMA 2016), Springer, 2016: 764-774. **(co-first author)**

9. **Y. Zheng**, C. Sun, C. Zhu, et al. LWCS: A large-scale web page classification system based on anchor graph hashing [C], 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2015: 90-94.

10. H. Peng, **Y. Zheng**, J. Li, et al. CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling [J]. Bioinformatics, 2018, 34 (18), 3069-3077.

11. H. Peng, **Y. Zheng**, J. Li et al. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions [J]. Bioinformatics, 2018, 34 (17), i757-i765.

12. H. Peng, C. Lan, **Y. Zheng**, J. Li, et al. Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite [J]. BMC bioinformatics, 2017, 18(1): 193.

13. X. Zhang, Z. Zhao, **Y. Zheng**, and J Li. Prediction of Taxi Destinations Using a Novel Data Embedding Method and Ensemble Learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2019.

14. X. Zhang, Y. Liu, **Y. Zheng**, Z. Zhao, J Li et al. Distinction Between Ships and Icebergs in SAR Images Using Ensemble Loss Trained Convolutional Neural Networks [C]. Australasian Joint Conference on Artificial Intelligence. Springer, 2018: 216-223.

15. Z. Zhao, H. Peng, C. Lan, **Y. Zheng**, J. Li et al. Imbalance learning for the prediction of N 6-Methylation sites in mRNAs [J]. BMC genomics 19 (1), 574.

Abstract

This research employs text and data mining methods to gain valuable knowledge for human drugs. Specifically, computational methods are developed for three topics, namely drug-side-effect prediction, drug-target identification, and drug-drug-interaction detection. The key innovations of the proposed methods lie in the feature space construction using medical domain knowledge, generation of reliable negative samples, and successful application of machine learning algorithms.

The drug-side-effect prediction problems are studied in Chapters 3-5. Side-effects are secondary phenotypic responses of human organisms to drug treatments. Side-effect prediction is an important topic for drugs especially in post-marketing surveillance because they cause significant fatality and severe morbidity. To overcome the limitations of existing computational methods such as lack of proper drug representation and reliable negative samples, this thesis presents three novel methods.

The first method is to predict side-effects for single drug medication as described in Chapter 3. A comprehensive drug similarity framework is developed by integrating several types of similarities measured by representative features of drugs first. Then reliable negative samples are generated through analyzing the comprehensive drug similarities. Trained with generated reliable negatives, the prediction performance of four classical classifiers are improved significantly, outperforming those state-of-the-art methods. Chapter 4 describes the method proposed to predict side-effects for combined medication of multi-drugs. A scoring method on a drug-disease-gene

tripartite network is developed to prioritize interacting drugs, paving a way to generate credible negative samples for side-effect prediction of combined medication. It creatively characterized a drug with its chemical structures, target proteins, substituents, and enriched pathways. The drug-drug pairs are represented as novel feature vectors to train binary classifiers for prediction. This novel representation and the inferred negative samples contribute to the superior performance of the proposed method in drug-drug-side-effect association prediction. Chapter 5 introduces the last method for detecting adverse drug reactions (ADRs, i.e., side-effects) from medical forums. It filters the cause-result relationship between drugs and ADRs using a self-built dictionary and detects drug-ADRs associations by information entropy. Compared with conventional co-occurrence based methods, the proposed method captures both high-frequency and low-frequency ADRs simultaneously. Besides, it returns drug-related ADRs only owing to the self-built relation dictionary.

Drug-target identification plays a crucial role in drug discovery. Existing computational methods have achieved remarkable prediction accuracy, however, usually obtain poor prediction efficiency due to computational problems. Chapter 6 presents a method to improve the prediction efficiency using an advanced technique named anchor graph hashing (AGH). AGH embeds data into low-dimensional Hamming space while maintaining the neighbourship. It turns the drug-target identification problem into a binary classification task where inputs are AGH-embedded vectors of drug-target pairs, and labels are judgments of their associations. Ensemble learning with random forest and XGBoost is employed to learn a good decision boundary. The proposed method is demonstrated to be the most efficient method and achieves comparable prediction accuracy with the best literature method.

Chapter 7 introduces a novel positive-unlabeled learning method named DDI-PULearn for large-scale detection of drug-drug interactions (DDIs). DDI-PULearn first generates seeds of reliable negatives via OCSVM (one-class support vector machine) under a high-recall constraint and via the

cosine-similarity based KNN (k-nearest neighbors) as well. Then trained with all the labeled positives (i.e., validated DDIs) and the generated seed negatives, DDI-PULearn employs an iterative SVM to identify the set of entire reliable negatives from the unlabeled samples. The identified negatives and validated positives are represented as vectors using the bit-wise similarity of corresponding drug pairs to train random forest for prediction. Its excellent performance is confirmed by comparing with two baseline methods and five state-of-the-art methods.